UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

AN INTELLIGENT MODEL FOR DETECTING FRAUD IN THE NON TECHNICAL LOSS OF COMMERCIAL POWER:

CASE OF KENYA POWER-RUIRU AREA.

BY

LELENGUIYA JOB KUNTAI

A RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE AT THE UNIVERSITY OF NAIROBI.

JULY, 2015

DECLARATION

This project, as presented in this report, is my original work and has not been presented for any other award in any other University.

Signature: _____ **Date:** _____

Name: Lelenguiya Job Kuntai

Reg. No: P52/69253/2013

This project report has been submitted as partial fulfillment of the requirements for the degree of Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University supervisor.

Signature: _____ **Date:** _____

Name: Dr. Elisha T. Opiyo.Omulo

DEDICATION

I dedicate this project report to my beloved mother Pauline Lelenguiya who despite having not gone through formal education, she knows its value and to my siblings Zak, Tom, Triza, Joyce and Josephine for their love and care they showed throughout my studies. You indeed have been of great help and blessings in every step of my life. May God bless you abundantly.

ACKNOWLEDGEMENT

First and foremost, I wish to take this opportunity to thank almighty God for the gift of life and for giving me strength and courage to complete this research on time. I also wish to express my sincere gratitude to all those who gave me encouragement and support towards the completion of this study. I am particularly thankful to my supervisor, Dr. Elisha T. Opiyo, for his patience and guidance in leading me through the research process, encouragement and support. I am deeply indebted to you for the help, suggestions and encouragement I received during the research and when writing this report. Finally to all my friends and family, Joseph Lesukat, Demalee, Pesh, Roselah, Nchamuni, colleagues and lecturers who have been involved in one way or the other in helping me move to this milestone, I say thank you this research project would not have been completed without you. May God reward you greatly.

ABSTRACT

Financial frauds are on the rise as a result large amount of money is lost by institutions to fraudsters, recognizing the problem of losses and the area of suspicious behavior is the challenge of fraud detection. Applying data mining techniques on financial statements can help in pointing out the fraudulent usage. It is important to understand the underlying business objectives to apply data mining objectives. Electricity consumer dishonesty is the main problem faced by power utilities that are managed by a financial billing system worldwide. Finding efficient measurements for detecting fraudulent electricity consumption has been an active research area in recent years. This research report presents a proposed model for detecting Non-Technical Loss (NTL) of commercial in electricity consumption utility using data mining techniques such as support vector machine, neural network, K-Nearest Neighbor and Naïve bayes. This work applies a suitable data mining technique in this field based on the customer information billing system for electricity consumption in selected accounts at Kenya Power Limited. The selected techniques are used in the design and development of a fraud detection model. The efficiency and accuracy of the model were tested and evaluated in order to get one accepted technique to be adopted by the Kenya Power Limited. From the results of the tested model, the biggest score for the fraud detection hitrate is achieved by support vector machine (SVM) classifier with 86.44% followed by K-Nearest Neighbor with 84.75% and classifier with the least optimal fraud detection rate is the Naïve bayes at 74.58%. Therefore this study adopted the SVM classifier for the following reasons. First, balancing technique used for ANN and K-Nearest Neighbor depend on random sampling in which decrease in the number of instances in the training data set to more than the half. Second, the SVM classifier depends on class weighting technique to balance data set without omitting any instance and finally SVM got the maximum accuracy score with balanced data sets.

Table of Contents

LIST OF FIGURES

LIST OF TABLE

LIST OF ABBREVIATIONS

AI- Artificial Intelligence

ANN- Artificial Neural Network

NTL- Non-Technical Losses

KNN- K-Nearest Neighbor

T&D-Transmission and Distribution

SVM- Support Vector Machine

SVC-Support Vector Classification

LV-low voltage

ERB- Electricity Regulatory Board

ERC- Energy Regulatory Commission

SNN-Simulated Neural Network

CIBS-Customer information billing system

AMR- Automatic Meter Reading

CHAPTER ONE

1.0 INTRODUCTION

1.1 Non-Technical Electricity Losses

Non-Technical Losses (NTLs) of power is the main problem being faced in the electricity supply industry. Such losses may occur due to meter tampering, meter malfunction, illegal connections, billing irregularities and unpaid bills. The problem of NTLs is not only faced by developing countries but also by developed countries such as China (6%), France (5%), Italy (7) and Japan at 5% as per the World Bank dataset (2009-2013).

Generation, transmission and distribution of electrical energy involve many operational losses. Whereas, losses concerned to generation can be technically defined, Transmission and Distribution (T&D) losses cannot be precisely quantified from the sending end information. This illustrates the involvement of some non technical parameters in T&D of electrical energy. Technical losses occur naturally as a result of power dissipation in transmission lines, transformers, and other power system components. Technical losses in T&D are computed from the information about total load on the grid and total energy billed (Oliveira and Barioni, 2009). Non-technical losses (NTL) are caused by actions external to the power system and consist primarily of electricity theft, non-payment by customers, and errors in accounting and record-keeping. NTL cannot be precisely computed, but can be estimated. So, it is very difficult to represent electricity theft as a function of actions responsible for the same. In many developing countries, NTL accounts to about 10– 40% of their total generation capacity (Smith, 2003).

Electricity theft forms a major chunk of the NTL. It includes illegal tapping of electricity from the feeder, by-passing the energy meter, tampering with the energy meter and several physical methods to evade payment to the utility company (Dick 1995). Of which, illegal tapping of electricity and tampering with energy meter are the most identified and accounted ways of theft. Electricity theft can also be defined as using electricity from utility company without a contract or valid obligation to alter its measurement (Smith, 2003). Electricity theft is a complex problem with many parameters to be evaluated before implementing any measures to detect and control that. These parameters include some issues like social, economical, regional, managerial,

political, infrastructural, literacy rate, criminal, mafia, corruption, effect on genuine customers, power quality, safety, and uncertainty in time period during which electricity is stolen.

Theft refers to the generic term for all crimes in which a person intentionally and fraudulently takes personal property of another without permission or consent and with the intent to convert it to the taker's use (including potential sale). Electricity theft is defined as a conscience attempt by a person to reduce or eliminate the amount of money he or she will owe the utility for electric energy. This could range from tampering with the meter to creating false consumption information used in billing to making unauthorized connections to the power grid. Metering and billing for electricity actually consumed by users is integral to commercial management of an electricity utility. Generation, transmission and distribution of electrical energy involve many operational losses (Smith, 2003).

The generation, transmission and distribution aspects of electricity provision are unbundled in Kenya. Non-technical losses are borne by the distribution firm. There are two dominant methods of electricity theft in Kenya, illegal hookups and meter tampering. Illegal hookups are mostly found in the inner city residents and are most easily detected. However, it is very difficult for enforcers to have these connections removed without the assistance of the police because of the resistant nature of these residents. There is a high risk to the culprits as well because of frequent electrical fires and electrocutions. The more wealthy residents and the commercial sector are the main culprits of meter related theft which can be meter tampering or bypassing the meter. This is a more sophisticated method of theft which is also more expensive.

This method of theft is harder to detect and can only be confirmed by a random audit. There is a potential for very large amounts of electricity generated being lost to theft because of the lower rates of detection and the high usage activities associated with meter related theft (KPLC, 2011). Kenya power currently supply electricity to about 2.1 million customers and non technical losses stands at 5% which result to loss of company's revenue of Kshs 2.2 billion annually.

The knowledge of electricity customers provides an understanding of their consumption behavior, which has recently become important in the electricity supply industry. With this knowledge, electricity providers are able to develop new marketing strategies and offer services based on customer demand. One of the most commonly method used in acquiring knowledge of

customers' behavior is load profiling, which is defined as "the pattern of electricity consumption of a customer or group of customers over a period". Load profiling has been used for many years by power utilities for tariff formulation, system planning, and devising marketing strategies.

Power utility companies record historical customer data, such as contractual details, billing procedures, and consumption records in various customer databases to support their billing activity. However, such information as it presently exists is often too complex to allow the human mind to formulate efficient and strategic decisions or draw effective conclusions. In addition, this information is often inaccessible and extremely time consuming to retrieve, due to the problems associated with archived data in complex database systems.

Due to the problem associated with NTLs in electric utilities, various methods for efficient management of NTLs and protecting revenue in the electricity distribution industry have been proposed. The most effective method currently to reduce NTLs and commercial losses up to date is by using smart and intelligent electronic meters, which make fraudulent activities more difficult, and easy to detect. However, the cost of such meters comes at an expensive price. Therefore, these types of meters are not currently feasible to be used by power utilities throughout the entire low voltage (LV) distribution network, i.e., in residential and commercial sectors.

In this research, data mining is employed to meet the above challenges. In general, data mining is defined as "a process of discovering various models, summaries and derived values from a given collection of data".

The main motivation of this study is to investigate the capability of using machine learning techniques such as Support Vector Machine (SVM), Neural Network, K-nearest neighbor and Naïve bayes for the detection and identification of NTL activities.

1.2 Statement of the problem

In the Kenya Power, kWh (Kilowatt-hour) meters record readings are used to bill customers for the amount of electricity consumed. Currently, electro-mechanical induction electricity meters are used in most of the homesteads, which can easily be tampered. Inspection of electric meters for fraud detection and identification is a manual and tedious task, which requires experienced staff and lot of time.

The actions taken by the Kenya Power in order to address the problem of NTLs include: meter checking and premise inspection, reporting on irregularities, and monitoring of unbilled accounts, which have resulted in fraud detection hit rate still remaining high. This is because customer installation inspections are carried out without any specific focus or direction. In most cases, inspections are carried out at random, while some targeted raids are undertaken based on information reported by the public or meter readers. With the application of an intelligent detection system using machine learning techniques, customer installation inspections are guided to a reduced group of likely fraudulent customers. Therefore, an intelligent model to guide the inspections of electricity meters is proposed in this research study.

1.3 The goal of the research

The main aim of the research was to develop an intelligent fraud detection model that will be able to help Kenya Power identify, detect and predict customers with fraudulent activities by investigating abrupt changes in customers' monthly load consumption patterns.

1.4 Objectives of the research

a) To develop an automated strategy to preprocess customer data for smoothening noise, feature extraction, data normalization and classification.
b) To develop a fraud detection model and compare their performances using Artificial Intelligence technique such as Support Vector Machine (SVM), Neural Network, K-Nearest neighbor and Naïve bayes.

1.5 Research Questions

a) What are strategies to preprocess customers' data for noise smoothening, features extractions and data normalization?

b) Which is the best classifier for detection of fraud in the non-technical loss of commercial power?

1.6 Significance of the Research

The developed fraud detection model in this research study will provide information to Kenya Power for efficient detection and classification of NTL activities in order to increase effectiveness of their onsite operation. With the implementation of the proposed model, operational costs due to onsite inspection in monitoring NTL activities will be significantly reduced. This will also reduce the number of inspections carried out at random, resulting in a higher fraud detection hitrate.

1.7 Scope of the research

The researcher focused on a small percentage of the Kenya Power customers low voltage usage residing at Ruiru area by using the monthly kWh interval data retrieved from the Kenya power database over a period of thirty six (36) months.

CHAPTER TWO

2.0 LITURATURE REVIEW

2.1 Kenya Power & Lighting Company Limited

The power industry in Kenya dates back in 1922 when East African Power and Lighting Company were incorporated, to generate and distribute electricity throughout the country. It changed its name to Kenya Power and Lighting Company limited (KPLC) through a special resolution by shareholders in 1983. Since then, the company has undergone several changes in response to the changing environment. For example the company no longer handles the generation of electricity, but only deals with transmission, distribution and retail of the electricity. The Kenya Electricity Generating Company (KENGEN) came into being in 1997, to focus on generation of electricity. This also opened the generation industry to independent power producers (IPPS) like Ibel Africa, Tsavo Power, among others. They all generate electricity and sell to KPLC in bulk at agreed prices. The Electricity Regulatory Board (ERB) was established by the electric power act, 1997, and was given the primary role of regulating the generation, transmission and distribution of electricity, promote and ensure competition in the sub-sector, approve contracts for generation and bulk sale of electricity and set or review electricity tariffs. The mandate was expanded and Energy Regulatory Commission (ERC) was established as an energy sector regulator in July 2007 under the Energy Act 2006. It is a single sector regulatory agency with responsibility for economic and technical regulation of electric power, renewable energy and downstream petroleum sub-sector, including tariff setting and reviewing, licensing, enforcement, dispute settlement and approval of power purchase and network service contracts.

In 2005, the rural electrification function was also hived from KPLC to form the Rural Electrification Authority (REA) which is responsible for construction of rural projects funded by existing consumers, government and donors. This would also improve connectivity of the rural areas. Despite the various changes in the power sector the company continued to make loses. In 2006, KPLC signed a management contract with Manitoba Hydro International (MHI), a Canadian power utility firm. The MHI was to offer management services to the company for two years between July 2006 and June 2008 and help the company to recover from the loss making regime to profit making. One of their terms of reference was to improve the system efficiency,

which is a measure of how much of the units purchased are finally wheeled to the customers, by 1.5% every year. In order to address this, the MHI management identified non-technical electricity losses as one of the areas to address. Various strategies were formulated to deal with non-technical electricity losses. This included intensified inspections of high risk areas and testing all large power customers' installations and recovering lost revenue through debiting the accounts with the lost revenue.

In 2008, Kenya Transmission Company (KETRACO) was formed to focus on construction of new transmission lines so that KPLC could focus more on service delivery. KPLC provides electricity energy that is used for various economic and social activities. Electricity consumption has traditionally been used as a reliable indicator of the economic activity and the standards of living in a given location, the increase in power consumption in a country was very significant. In the year July 2009 to June 2010, the sales of electricity grew by 8% to indicate that the society has adapted use of electricity as a way of life (KPLC annual reports, 2010). However, there has not been significant improvement in the system efficiency despite the various efforts put in place by the organization through system reinforcements. This can be attributed to increase in non-technical electricity losses as the economy declines and people's purchasing power is depleted due to inflation.

Due to the ever changing environment, KPLC has developed a 5 year strategic plan which includes losses reduction as a strategy in various business level activities (KPLC, 2011). Due to many developments and changing customer demands, it has been necessary for KPLC to carry out a rigorous rebranding exercise. The corporate rebranding and organizational culture change project (Project Mwangaza) commenced in January 2010 with a series of staff workshops that developed a new set of Vision, Mission and Values. Training on high performance leadership has been conducted for several members of staff. This activities have culminated in the unveiling of new company logo in June 2011 and the company is now referred to as Kenya Power as a new brand name. The company has also adapted mission leadership as a way of reviewing business processes and measuring the success of the activities being carried out on a monthly basis. It is expected that all these strategies will help the organization to achieve its corporate objectives more successfully.

2.2 Data Mining

The benefits of data mining can be extracted from knowing its definition. Data Mining is the science of extracting useful information from large datasets or database is known as data mining. It is a new discipline, lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas (AihuaShen and Rencheng Tong 2007). The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data need to be converted into information and knowledge from recorded data to become useful. Traditionally, analysts have performed the task of extracting useful information from the recorded data, but the increasing volume of data in modern business and science calls for computer-based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. Data mining techniques play a big role in analyzing large databases with millions of records to achieve new unknown patterns. These patterns help in the study of customer transactions behavior in many scientific researches as in mine. When we get the suspicious customer patterns using data mining techniques, then we can narrow the circle of investigation to get fraudulent issues very fast according to the manual way that could be impossible.

2.3. Types of Data Mining.

Data mining techniques are divided into two kinds as supervised and non supervised techniques. “In supervised modeling, whether for the prediction of an event or for a continuous numeric outcome, the availability of a training dataset with historical data is required. Models learnt from past cases” (Konstantinos Tsiptsis and Antonios Chorianopoulos, 2009). Supervised techniques such as classification models depend on a label class which must exist in the database under study. This type of mining is used in several fields like direct marketing, Credit/loan approval, Medical diagnosis if a tumor is cancerous or benign**,** fraud detection if a transaction is fraudulent, etc.In the non-supervised learning such as clustering models, there is no class label. Clustering is a method of grouping data that share similar trend and patterns, applied in Insurance, city-planning to identify groups of houses according to their house type, value, and geographical location.

2.3.1 Data Mining Methodology

The data mining methodology consists of the following steps.

1. Data cleaning to remove noise and inconsistent data.
2. Data integration where multiple data sources may be combined.
3. Data selection where data relevant to the analysis task are retrieved from the database.
4. Data transformation where data are transformed into forms appropriate for mining by performing summary. Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
5. Pattern evaluation to identify the truly interesting patterns representing knowledge based on some interesting measures.
6. Knowledge presentation where visualization and knowledge representation techniques are used to present the mined knowledge to the user

2.4 Classification

Classification models are linear and non-linear. "Linear models are analytical models that assume linear relationships among the coefficients of the variables being studied." (webdocs.cs.ualberta.ca, 2012). Furthermore, defined as "Approximation of a discriminant function or regression function using a hyper plane. It can be globally optimized using simple techniques, but does not adequately model many real-world problems". The other type of classification prediction models is the non-linear models defined as "Non-Linear Predictive Models are analytical model that does not assume linear relationships among the coefficients of the variables being studied." (statsoft.com, 2014). Linear models in spite of their computational simplicity, stability they have an obvious potential weakness. "The actual process may not be linear, and such an assumption introduces uncorrectable bias into the predictions."(wikipedia.org, 2014).

2.4.1 Support Vector Machines

The first selected mining technique is SVM that can be used as linear and non linear. "In machine learning, SVM is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. Moreover a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space which can be used for classification regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mapping used by the SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function K(x,y) selected to suit the problem. The hyperplanes in the higher dimensional space are defined as the set of points whose inner product with a vector in that space is constant in addition to performing linear classification. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces." (Rapid miner user manual, 2010).

2.1.2 The Case When The Data Are Linearly Separable

SVMs are supervised learning methods that generate input-output mapping functions from a set of labeled training data. “The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output)” (Han and Kamber, 2005). For classification, “nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data” (Han and Kamber, 2005). Two parallel hyperplanes are constructed on each side of the hyperplanes that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be (Han and Kamber, 2005). “SVMs can be used for prediction as well as classification. They have been applied to a number of areas including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests” (Maimon and LiorRokach, 2008). SVMs have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, Bioinformatics, face recognition, image processing and text mining, which has established SVMs as one of the most popular, state-of-the-art tools for knowledge discovery and data mining. Similar to artificial neural networks, SVMs possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. Therefore, they are of particular interest to modeling highly nonlinear, complex systems and processes. (Han and Kamber, 2005).

2.3.3 The Case When The Data Are Linearly Inseparable

The linear SVMs would not be able to find a feasible solution here. "Now what the good news is that the approach described for linear SVMs can be extended to create non-linear SVMs for the classification of linearly inseparable data (also called non-linearly separable data, or nonlinear data, for short). Such SVMs are capable of finding non-linear decision boundaries (i.e., nonlinear hyper surfaces) in input space." So, the question is how to extend the linear approach? Obtaining a nonlinear SVM can be done by extending the approach for linear SVMs as follows. There are two main steps. In the first step, transform the original input data into a higher dimensional space

using a non-linear mapping. Once the data have been transformed into the new higher space, the second step searches for a linear separating hyperplane in the new space. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hyper surface in the original space. The complexity of the learned classifier is characterized by the number of support vectors rather than the dimensionality of the data. Hence, SVMs tend to be less prone to overfitting than some other methods (Maimon and LiorRokach, 2008). "The concept of overfitting is very important in data mining. It refers to the situation in which the induction algorithm generates a classifier which perfectly fit the training data but has lost the capabilities of generalizing to instances not presented during training. In other word, instead of learning, the classifier just memorizes the training instances". (Myatt, 2007).

2.5 Neural Network

The second selected mining technique is ANN which is applicable when working with the non-linear problems as done in (Nagi, 2010). This technique used successfully in business fraud detection, and defined as "artificial neural networks (ANNs) are widely used for fraud detection" (Nagi, 2010), "a neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms, neural networks are non-linear statistical data modeling or decision making tools" (Rapid miner user manual, 2010). Furthermore, "The inspiration for neural networks was the recognition that complex learning systems in animal brains consisted of closely interconnected sets of neurons. Although a particular neuron may be relatively simple in structure, dense networks of interconnected neurons could perform complicated learning tasks such as classification and pattern recognition" (Rapid miner user manual, 2010). The topologies of neural networks or neural network architectures formed by organizing nodes into layers and linking these layers of neurons with modifiable weighted interconnections. In recent years, neural network researchers have incorporated methods from statistics and numerical analysis into their networks. Being a nonlinear mapping relation from the input space to output space, neural

networks can learn from the given cases and summarize the internal principles of data even without knowing the potential data principles ahead. It can adapt its own behavior to the new environment with the results of formation of general capability of evolution from present situation to the new environment (Rapid miner user manual, 2010).

From the aspect of the pure theory, the nonlinear neural networks method is superior to the statistical methods in the application for credit card fraud detection such as in (AihuaShen and Rencheng, 2007). It is sometime unusual in the practice research even though the common advantages of the neural networks as a possible result of usage of improper network structure and learning computing method. On the other hand, there are still many disadvantages of the neural networks, such as the difficulty to confirm the structure, the efficiency of training, excessive training, and so on. The researchers in (AihuaShen and Rencheng, 2007) conduct a comparison between three classification methods were tested for their applicability in fraud detection, i.e. decision tree, neural networks and logistic regression. The three methods are compared in terms of their predictive accuracy. Neural network classifier has the best accuracy in testing results.

2.6 K-Nearest-Neighbor

The last selected mining method is KNN, which also supports non-linear problem. This section reviews some point about KNN. "The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition" (Han and Kamber, 2005). "In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally, and all computations is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor." The Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attribute. Each tuple represents

a point in a n-dimensional space. In this way, all of the training tuples are stored in a n-dimensional pattern space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k nearest neighbors of the unknown tuple. “There are many methods for determining whether two observations are similar. For example, the Euclidean or the Jaccard distance. Prior to calculating the similarity, it is important to normalize the variables to a common range so that no variables are considered to be more important” (Rapid miner user manual, 2010). “Closeness” is defined in terms of a distance metric, such as euclidean distance.

2.7 Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be 'independent feature model'. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix.

2.8 Measuring Performance

After model building, knowing the power of model prediction on a new instance, is very important issue. Once a predictive model is developed using the historical data, one would be curious as to how the model will perform on the data that it has not seen during the model building process. One might even try multiple model types for the same prediction problem, and then, would like to know which model is the one to use for the real-world decision making situation, simply by comparing them on their prediction performance (e.g., accuracy). To measure the performance of a predictor, there are commonly used performance metrics, such as

accuracy, recall etc. First, the most commonly used performance metrics will be described, then some famous estimation methodologies are explained and compared to each other. "Performance Metrics for Predictive Modeling In classification problems, the primary source of performance measurements is a coincidence matrix (classification matrix or a contingency table)" (Han and Kamber, 2005). Figure 2.7 shows a coincidence matrix for a two-class classification problem.

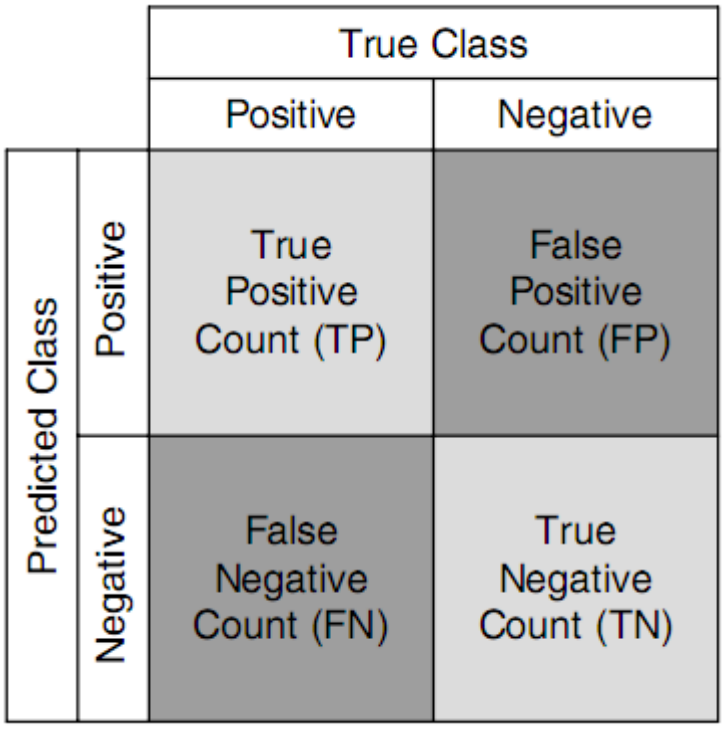


Figure 2.7 a simple confusion matrix as per (Han and Kamber, 2005)

2.8.1 The K-Fold Cross Validation

When the amount of data for training and testing is limited, t holdout method reserves a certain amount for testing and uses the remainder for training. In practical terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training. Of course, the selected third for testing or training may be unlucky: The sample used for training or testing might not be representative. "In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called k-fold cross validation. In k-fold cross validation,

also called rotation estimation, the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. The classification model is trained and tested k times. Each time it is trained on all but one folds and tested on the remaining single fold" (Han and Kamber, 2005).The cross validation overall accuracy model calculated by simply averaging the k individual accuracy measures.

2.9 Non Technical Loss Detection For Metered Customers In Power Utility

In (Dianmin Yue and Xiaodan Wu,2007) the researchers design and develop a fraud detection system for electricity metered customers, the data were collected from the power utility in kula-lambour, after feature extraction and selection, they define the normal load profile dataset and the fraudulent load profile dataset representing them as in Fig 2.8 shown

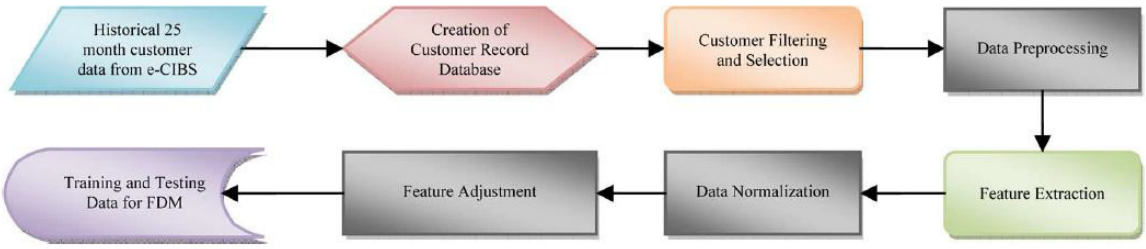


Figure 2.8 Proposed framework for processing e-CIBS data for C-SVM training and validation (Dianmin Yue and Xiaodan Wu, 2007).

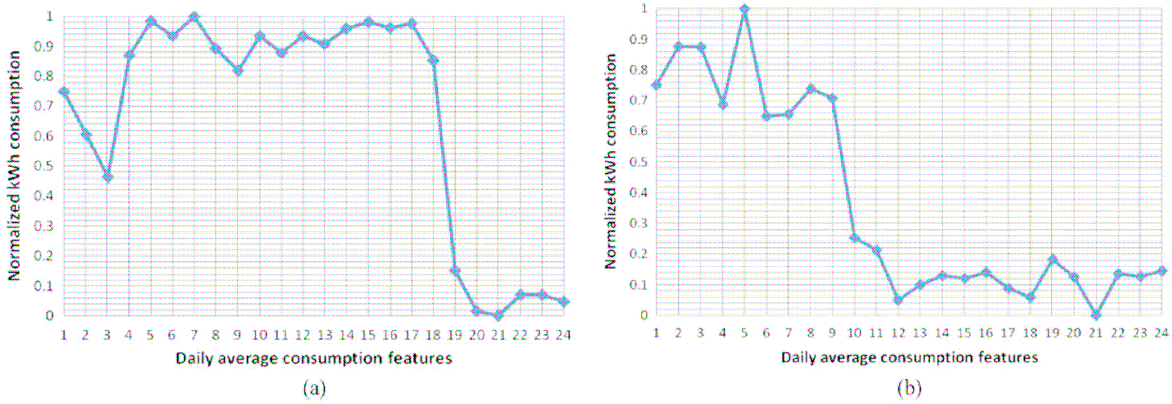


Figure 2.8.1 Normalized load profiles of two typical fraud customers over a period of two years**.**

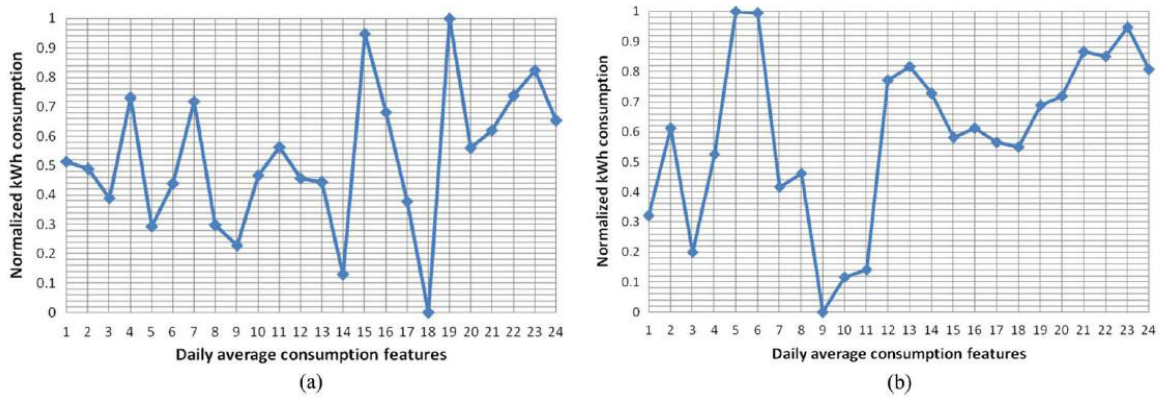


Figure 2.8.2. Normalized load profiles of two normal customers over a period of two years. (Dianmin Yue and Xiaodan Wu, 2007).

By applying the SVM classification method, they improve the hit rate in fraud detection from the manual detection by the fraud team with 3% hit rate to 60% by the SVM model.

2.9.1Financial Fraud Detection

In (Chang and C.-J. Lin, 2003) the researchers from China and U.S.A analyzed and studied the used data mining techniques in 18 firms as a study reference, then they proposed a generic framework for data mining based on financial fraud detection as in Figure 2.9.

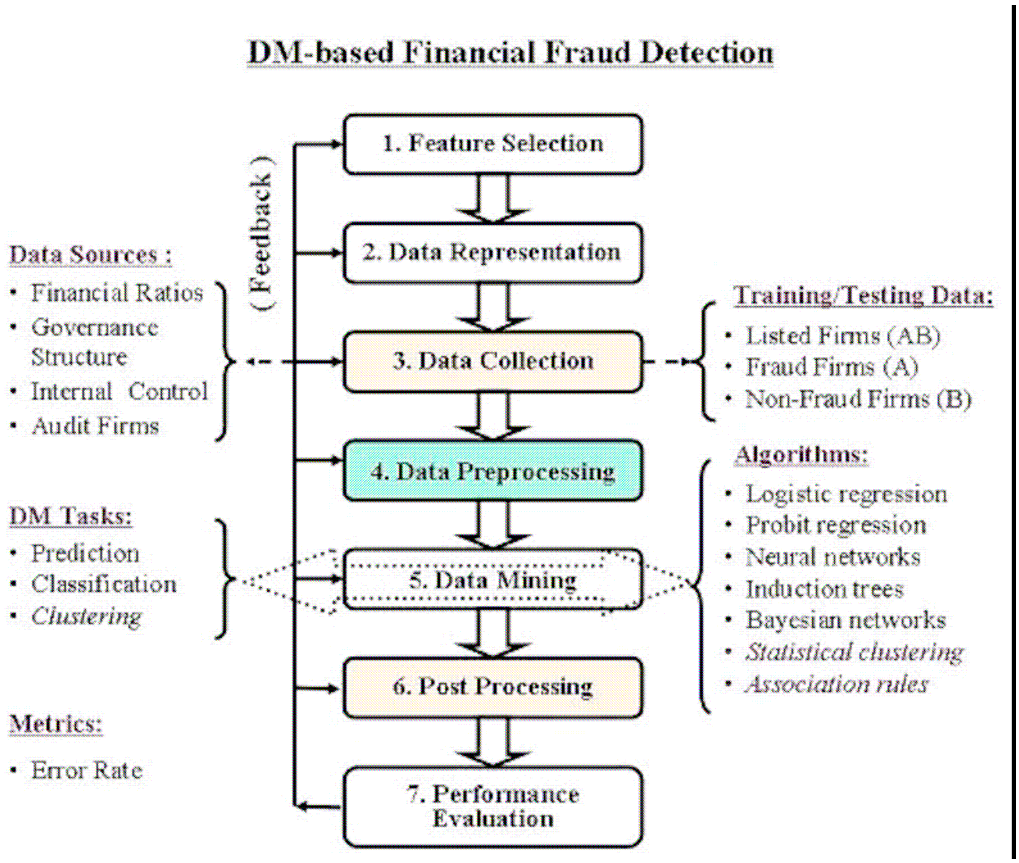


Figure 2.9.1 A Generic framework for DM-based FFD as in (Chang and C.-J. Lin, 2003) It is clear that classification methods are the dominant techniques in this generic frame work.

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Introduction

This section covered the methodology of the research used by the researcher. It is comprised of the following; research design, data collection, conceptual framework and data analysis.

3.2 Research Design

The research design for this study was a case study and focused on the Kenya Power Ltd. This study aimed at collecting information from respondents on the causes of non-technical electricity losses in the Kenya Power with the view of proposing a suitable fraud detection model to address this challenge.

Research design refers to the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in the procedure (Babbie, 2002). In addition Kothari (2004) observed that research design is a blue print which facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible hence yielding maximum information with minimal expenditure of effort, time and money.

The design is deemed appropriate in this study because there's only one organization currently engaged in the distribution of electricity in Kenya. The research design helped the researcher to draw conclusions of the relationship between non- technical electricity losses and the strategic responses by Kenya Power.

3.3 Data Collection

The source of data was both primary and secondary. The instrument used was an in-depth personal interview guide. The questions were geared to acquire the opinion of the respondent on strategies by Kenya Power in reducing non- technical loss of electricity. The study considered six (6) respondents drawn from customer care and fraud control department of the Kenya power for an interview: The respondents were expected to give an insight into the strategic control practices in the respective positions. Secondary data was obtained from existing company records and customers information billing database for Kenya Power customers residing at Ruiru area. In this research study, the researcher used historical customer billing data of Kenya Power customers residing at Ruiru and transforms the data into the required format for the classifier, by data preprocessing and feature extraction. Ruiru customers are represented by their consumption profiles over a period of time. These profiles are characterized by means of patterns, which significantly represent their general behavior, and it is possible to evaluate the similarity measure between each customer and their consumption patterns. This creates a global similarity measure between normal and fraud customers as a whole.

The electricity consumption data that was used in this project and research study was identified by the fraud control department team experts who have inspected customer premises and meter installations giving fraud cases and by billing system experts who gave the historical customer electricity usage and billing information. The historical customer billings and consumption data were collected to train the proposed intelligent model to learn and differentiate between normal and suspicious consumption patterns.

The researcher was able to get electricity monthly consumption for 36 months for 963 customers representing consumption from July 2011 to July 2014. The data was obtained in the oracle database format and converted to excel files for data processing. Two types of customer data were collected, which are: The customer information billing system (consumption profiles) and the customer's electricity irregularities data (anomaly case).

3.4 Conceptual framework

The researcher adopted the CRISP-DM reference model shown in the fig 3.4 below for the design and devolvement of fraud detection model.

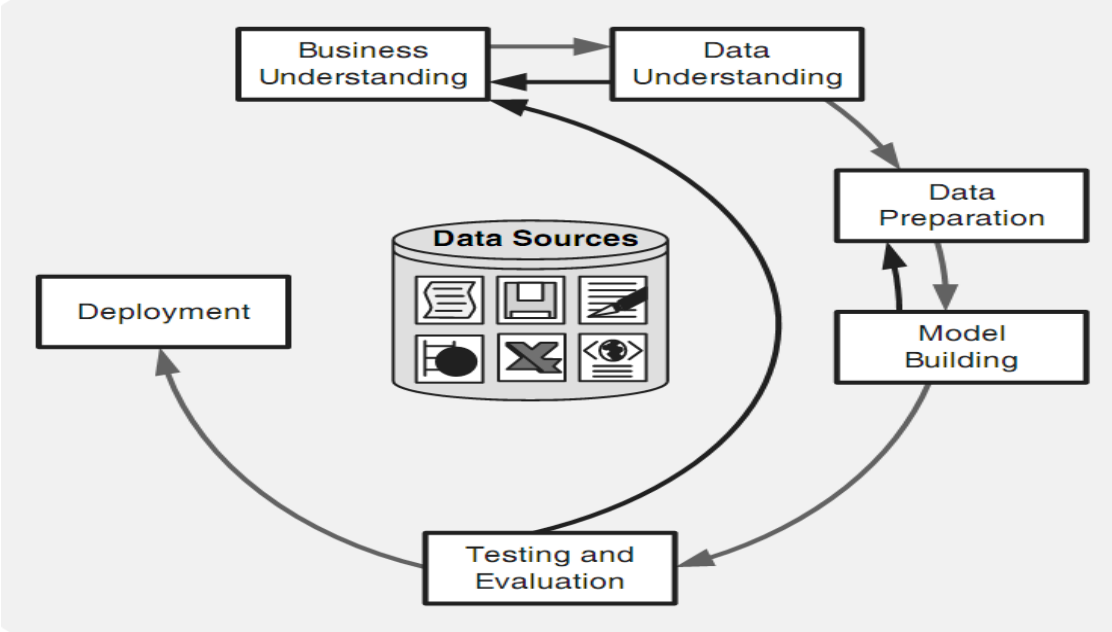


Figure 3.4 Phases of CRISP-DM reference mode.

CRISP-DM breaks the process of data mining into six major phases. The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The phases are illustrated below.

3.4.1 Business Understanding

In this initial phase, the researcher focused on understanding the project objectives and requirements from the Kenya power customer's electricity consumption pattern and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

3.4.2 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3.4.3 Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

3.4.4 Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

3.4.5 Evaluation

At this stage in the project you have built a model (or models) that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

3.4.6 Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases it will be the customer, not the data

analyst, who will carry out the deployment steps. Even if the analyst deploys the model it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

3.5 Data Analysis

Data was analyzed and evaluated using content analysis. The data collected was summarized according to the study objectives being causes of non- technical losses of commercial power at Kenya Power and data preparation for preprocessing was analyzed using Microsoft Excel and rapid miner studio 5.

CHAPTER FOUR

4.0 MODEL DESIGN AND DEVELOPMENT

The researcher adopted the methodology shown in figure 4.2 (Jawad Nagi, 2009) to design and develop an intelligence model for detecting and predicting fraud in the non-technical loss of commercial power at Kenya power.

In this research study, the NTL detection framework proposed in Figure 4.2 uses historical customer billing data of Kenya power customers' and transforms the data into the required format for the Support Vector Machine , Neural Network, K-nearest neighbor and the Naïve bayes by data preprocessing and feature extraction. Kenya power customers are represented by their consumption profiles over a period of time. These profiles are characterized by means of patterns, which significantly represent their general behavior and it is possible to evaluate the similarity measure between each customer and their consumption patterns. This creates a global similarity measure between a normal and fraud customers as a whole. The identification, detection and prediction are undertaken by the Support Vector Classification, which is the intelligent "classification engine".

The intelligent fraud detection model in this project was developed using rapid miner studio 5 software and Microsoft office excel 2007. The computer used for training was hp core i3 with 4.0GB RAM capacity.

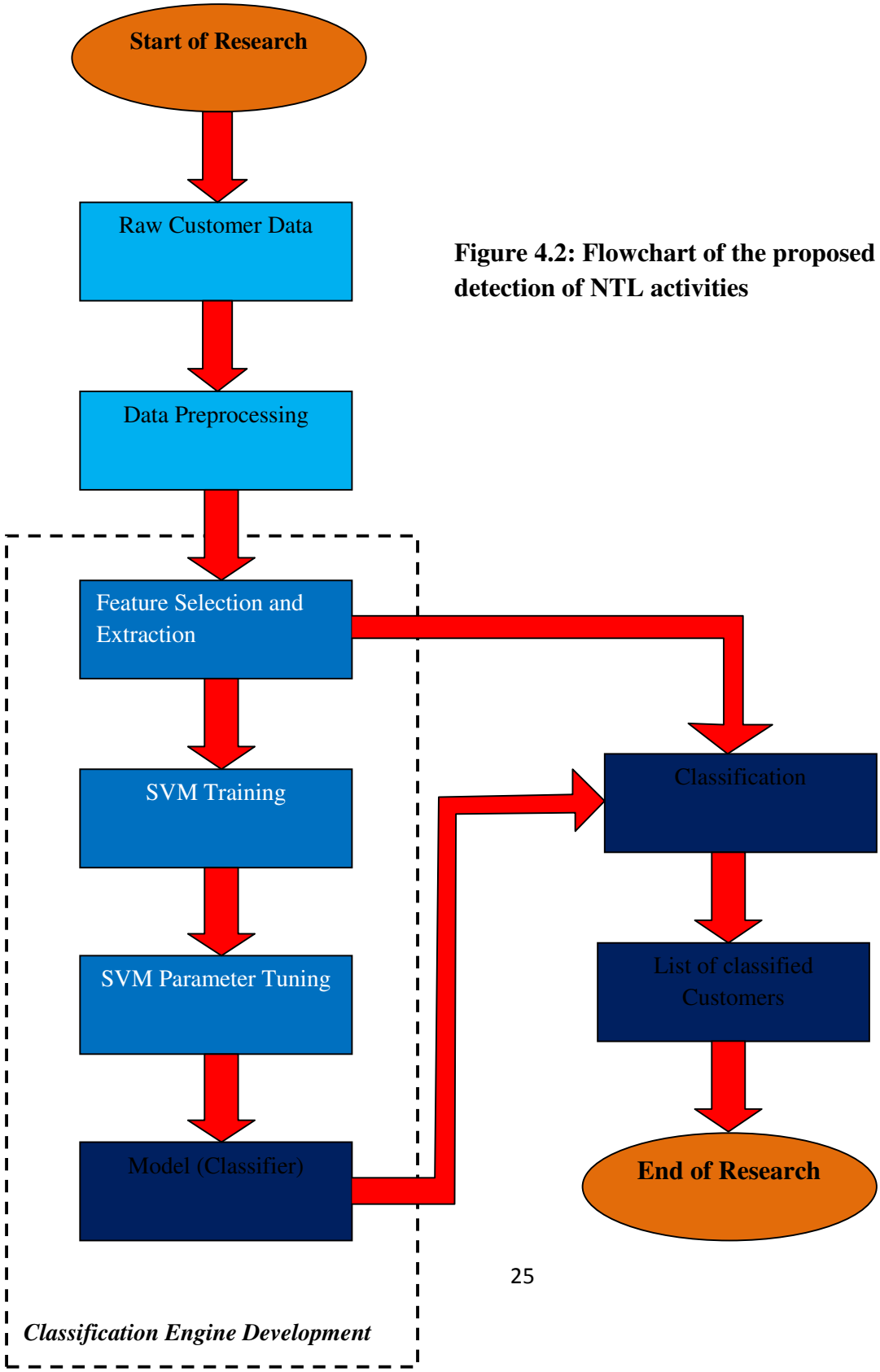


Figure 4.2: Flowchart of the proposed framework for detection of NTL activities

The data collected for this research was historical customer billing and consumption data for training the classifiers, i.e., to make the intelligent system learn, memorize and differentiate between normal and suspicious consumption patterns.

Kenya power customer's electricity consumption data for Ruiru residential area was obtained from oracle database and converted into excel data. Two types of customers were collected, normal customers and customers previously found with anomaly cases by the Kenya power inspection team.

4.1Customer Information Billing System Data

The customer information and billing data based on Oracle database was collected from the Kenya power customer care center. Two tables were extracted: customers with normal consumptions patterns and the one with abnormal consumption pattern. The following attributes were extracted from the historical monthly consumption table.

Column	Description
Meter no	Customer electricity meter number
Area	Customers location area
location	Geographical location of the meter
Meter status	To record if the meter is connected or disconnected
Monthly consumption	Monthly consumption
Reading date	Electricity consumption reading date

Table 3.1.1 customers table attributes for normal customers

4.2 Customers with anomaly

Customers found by the fraud control team with anomalies in their electricity meter for the 36 months were 59 in number. The table consisted of the following attributes.

Column	Description
Meter no	Customer electricity meter number
location	Geographical location of the meter
Breach date	The date of electricity breach

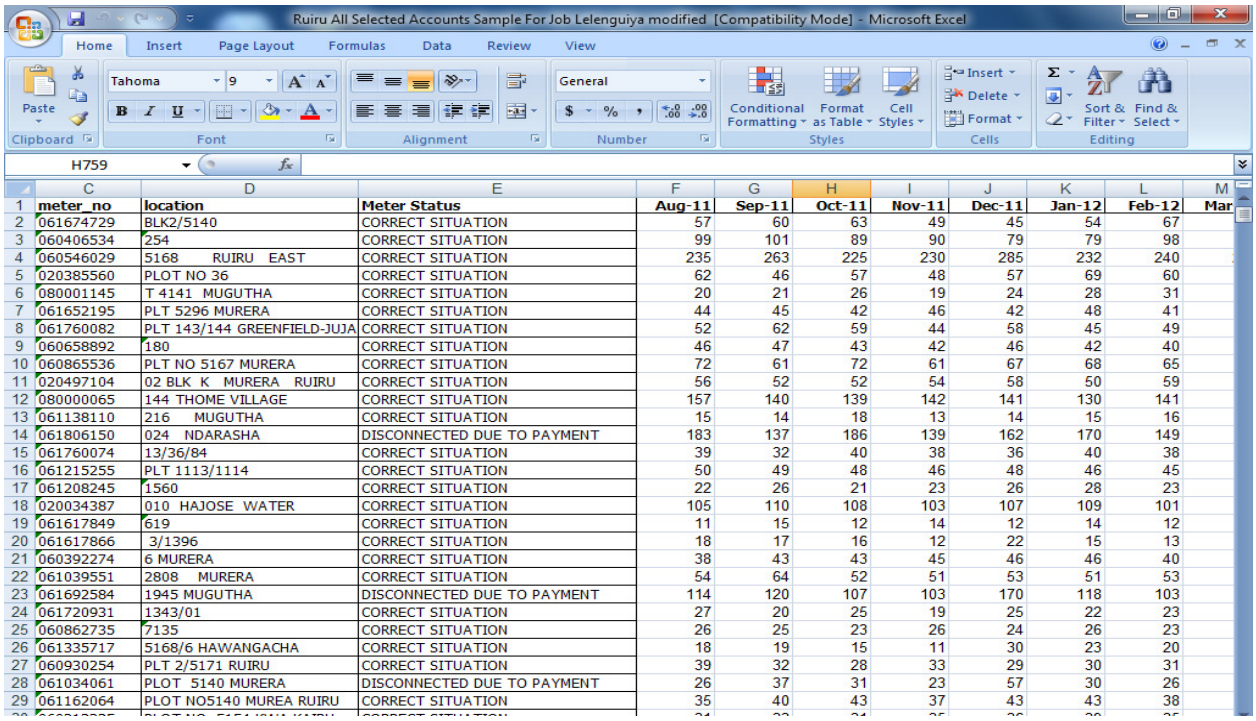
Table 3.1.2 customers table attributes **for disconnected customers**

4.3 Data Preprocessing

Data preprocessing is the first major stage involved in the development of the fraud detection system. Data preprocessing involves data mining techniques in order to transform raw customer data into the required format, to be used by the classifier for detection and identification of fraud consumption patterns.

4.3.1 CIBS and Customers with anomaly Data Preprocessing

The figure 3.1 shows monthly customer consumption for thirty six (36) months (July 2011 to July 2014) after filling consumption columns.



Ruiru All Selected Accounts Sample For Job Lelenguiya modified [Compatibility Mode] - Microsoft Excel

	C	D	E	F	G	H	I	J	K	L	M
1	meter_no	location	Meter Status	Aug-11	Sep-11	Oct-11	Nov-11	Dec-11	Jan-12	Feb-12	Mar
2	061674729	BLK2/5140	CORRECT SITUATION	57	60	63	49	45	54	67	
3	060406534	254	CORRECT SITUATION	99	101	89	90	79	79	98	
4	060546029	5168 RUIRU EAST	CORRECT SITUATION	235	263	225	230	285	232	240	
5	020385560	PLOT NO 36	CORRECT SITUATION	62	46	57	48	57	69	60	
6	080001145	T 4141 MUGUTHA	CORRECT SITUATION	20	21	26	19	24	28	31	
7	061652195	PLT 5296 MURERA	CORRECT SITUATION	44	45	42	46	42	48	41	
8	061760082	PLT 143/144 GREENFIELD-JUJA	CORRECT SITUATION	52	62	59	44	58	45	49	
9	060658892	180	CORRECT SITUATION	46	47	43	42	46	42	40	
10	060865536	PLT NO 5167 MURERA	CORRECT SITUATION	72	61	72	61	67	68	65	
11	020497104	02 BLK K MURERA RUIRU	CORRECT SITUATION	56	52	52	54	58	50	59	
12	080000065	144 THOME VILLAGE	CORRECT SITUATION	157	140	139	142	141	130	141	
13	061138110	216 MUGUTHA	CORRECT SITUATION	15	14	18	13	14	15	16	
14	061806150	024 NDARASHA	DISCONNECTED DUE TO PAYMENT	183	137	186	139	162	170	149	
15	061760074	13/36/84	CORRECT SITUATION	39	32	40	38	36	40	38	
16	061215255	PLT 1113/1114	CORRECT SITUATION	50	49	48	46	48	46	45	
17	061208245	1560	CORRECT SITUATION	22	26	21	23	26	28	23	
18	020034387	010 HAJOSE WATER	CORRECT SITUATION	105	110	108	103	107	109	101	
19	061617849	619	CORRECT SITUATION	11	15	12	14	12	14	12	
20	061617866	3/1396	CORRECT SITUATION	18	17	16	12	22	15	13	
21	060392274	6 MURERA	CORRECT SITUATION	38	43	43	45	46	46	40	
22	061039551	2808 MURERA	CORRECT SITUATION	54	64	52	51	53	51	53	
23	061692584	1945 MUGUTHA	DISCONNECTED DUE TO PAYMENT	114	120	107	103	170	118	103	
24	061720931	1343/01	CORRECT SITUATION	27	20	25	19	25	22	23	
25	060862735	7135	CORRECT SITUATION	26	25	23	26	24	26	23	
26	061335717	5168/6 HAWANGACHA	CORRECT SITUATION	18	19	15	11	30	23	20	
27	060930254	PLT 2/5171 RUIRU	CORRECT SITUATION	39	32	28	33	29	30	31	
28	061034061	PLOT 5140 MURERA	DISCONNECTED DUE TO PAYMENT	26	37	31	23	57	30	26	
29	061162064	PLOT NO5140 MUREA RUIRU	CORRECT SITUATION	35	40	43	37	43	43	38	

The figure 4.3.1 all Ruiru selected accounts.

The meter number indicates the account number of the customer. As customer data is confidential to the utility so in order to protect the privacy of the customers, "Customer Names" have been omitted. The two data sets Customer **Information Billing System** and Customers whose accounts have been disconnected are combined into one dataset with all attribute in the two datasets. Another new calculated attribute titled "CLASS" has been added in which each customer founded in disconnected accounts labeled with 'YES', the rest of other customers take "NO" value. This combined dataset was preprocessed to remove unwanted customers and smoothen out noise and other inconsistencies, in order to extract only useful and relevant information required as shown in the following three major steps that were involved in the preprocessing Customer Information Billing System data and data showing disconnected accounts. Customer Filtering and Selection, Consumption Transformation, Feature Selection and Extraction, Feature Normalization and classification.

4.3.2 Customer Filtering and Selection

Data from Kenya power was acquired in a raw format which consisted of 968 accounts with 85 accounts showing abnormal electricity consumption. In order to extract relevant and functional information, only customers with complete and useful data were selected from Ruiru residential area for the classification model development.

Since, the data acquired is in the form of a database, the Structured Query Language (SQL) applied to satisfy the criteria as listed:

a. Remove repeating customers in the monthly Customer Information Billing System data.
b. Remove customers having no consumption throughout the entire 36 months period.

After performing filtering and data reduction on the Ruiru data, only 702 customer records remained.

4.3.3 Consumption Transformation

Real world datasets tend to be noisy and inconsistent. This problem is overcome by data mining techniques using statistical methods were applied to the customer information billing system data, in order to remove noise and other inconsistencies. In some cases where meter readers are unable to record meter readings, due to customers not being present at their premises, therefore, in these situations meter readers indicate an "estimated" monthly kWh consumption for those customers in the billing information. This estimated billing consumption is based on the previous consumption recorded for the customers. Since, the estimated monthly consumptions are not an accurate consumption value, therefore, they were transformed into "normal" consumption values in order to smoothen out inconsistencies within the consumption patterns.

The estimated consumption transformation is accomplished by averaging all the previously consecutive normal "N" consumption values of a customer with respect to the currently estimated consumption. As an example, for an estimated consumption value, if the last two consecutive consumptions of a customer are normal, i.e., say N1 = 200 kWh and N2 = 300 kWh, then the average of N1 and N2 will be 250 kWh. This value of 250 kWh will be replaced in the estimated value as the normal (transformed) consumption value. This technique is implemented for the entire 36 month period, for all customers.

4.3.4 Feature Selection and Extraction

Particular features were selected from the Customer Information Billing System data, in order to extract only useful and relevant information required for training the classification model. Since the proposed fraud detection approach applies consumption information of customers (load profiles) to the detection of anomaly activities, therefore, the consumptions features are crucial part for pattern recognition in this research study. Other features were evaluated for selection based on feature selection using information gain method, which is a popular method used to reduce the dimensionality (Daniel Morariu, Lucian N. Vintan, and Volker Tresp, 2006). From table 4.2 three features which are considered to be useful to the problem of NTL detection were evaluated and appeared in the top of evaluation: Meter no, Meter status and location.

4.3.5 Feature Normalization

The selected features have a different scale, so, in order for the feature data to fit the SVM classification model properly all feature data were represented using a normalized scale. All data was linearly scaled using the z-transformation. Figure 4.3.5 represents the normalized features in columns and the customers indicated by the rows.

ExampleSet (702 examples, 1 special attribute, 37 regular attributes) View Filter (702 / 702): all

Row No.	Fraud_Status	meter_no	Aug-11	Sep-11	Oct-11	Nov-11	Dec-11	Jan-12	Feb-12	Mar-12	Apr-12	May-12	Jun-12	
1	No	0.404	0.017	0.104	0.202	-0.116	-0.275	-0.048	0.281	0.309	0.405	0.041	-0.235	0.4
2	No	0.316	1.119	1.207	0.877	1.057	0.643	0.594	1.087	0.364	1.157	0.969	1.044	1.0
3	No	0.326	4.686	5.561	4.408	5.062	6.202	4.522	4.779	4.362	4.636	5.961	5.915	5.6
4	No	-2.456	0.148	-0.272	0.047	-0.144	0.049	0.337	0.099	-0.080	0.249	-0.278	-0.144	0.0
5	No	1.674	-0.953	-0.944	-0.758	-0.974	-0.841	-0.716	-0.655	-0.830	-0.789	-1.004	-0.783	-0.
6	No	0.403	-0.324	-0.299	-0.343	-0.201	-0.356	-0.202	-0.395	-0.385	-0.400	-0.307	-0.205	-0.
7	No	0.410	-0.114	0.158	0.099	-0.259	0.076	-0.279	-0.187	-0.080	-0.296	0.041	0.130	-0.
8	No	0.334	-0.271	-0.245	-0.317	-0.316	-0.248	-0.356	-0.421	-0.330	-0.218	-0.278	-0.326	-0.
9	No	0.348	0.411	0.131	0.436	0.228	0.319	0.311	0.229	0.253	0.353	0.360	0.282	0.3
10	No	-2.448	-0.009	-0.111	-0.083	0.027	0.076	-0.151	0.073	-0.080	-0.244	-0.249	0.130	0.0
11	No	1.674	2.640	2.255	2.175	2.545	2.316	1.903	2.205	2.641	2.533	2.333	2.718	2.9
12	No	0.367	-1.084	-1.132	-0.966	-1.145	-1.111	-1.050	-1.045	-1.024	-0.945	-1.091	-1.118	-1.
13	No	0.410	-0.455	-0.648	-0.395	-0.430	-0.518	-0.408	-0.473	-0.580	-0.348	-0.337	-0.387	-0.
14	No	0.372	-0.166	-0.191	-0.187	-0.201	-0.194	-0.254	-0.291	-0.052	-0.114	-0.249	-0.205	0.0
15	No	0.372	-0.901	-0.810	-0.888	-0.859	-0.787	-0.716	-0.863	-0.719	-0.789	-0.801	-0.996	-0.
16	No	-2.480	1.276	1.448	1.371	1.429	1.398	1.364	1.165	1.308	1.547	1.579	1.653	1.3
17	No	0.400	-1.189	-1.105	-1.122	-1.117	-1.165	-1.075	-1.149	-1.218	-1.075	-1.207	-1.118	-1.
18	No	0.400	-1.006	-1.052	-1.018	-1.174	-0.895	-1.050	-1.123	-1.052	-1.049	-1.091	-1.148	-1.

Figure 4.3.5: The normalized features over a period of 36 months.

4.4 Classification Engine Development

The development of the classification engine, the SVM model, is the main focus of this project and research study. Development of the classification engine involves: load profile inspection for detection of normal and abnormal customers, training and development of the SVM classifier, SVM parameter tuning and SVM training and testing.

4.4.1 Load Profile Inspection

Load profiles i.e thirty six (36) months electricity consumption features of the customers were inspected to retrieve samples, in order to build the SVM model for the purpose of training. As in this study, a 2-class SVM classifier is used to represent two different types of customer load profiles, therefore, load profile samples used to build the SVM classifier were extracted from the

preprocessed Ruiru data. The load profiles inspected were extracted and classified into two different categories according to their behavior, i.e., fraud or normal consumption (non-fraud).
From the 702 filtered customers in the Ruiru data, only 59 customers were identified and detected as anomalous in the past 36 months. Customers with no fraud activities and the customers with anomalous activities form the backbone for the development of the SVM model.
Manual inspection was performed on a large sample of the remaining 643 customers. Customers' load profiles have been inspected and filtered in which abrupt changes appear clearly, indicating fraud activities, abnormalities and other irregularities in consumption characteristics. The 643 customers' load profiles selected as normal consumption in order to train the SVM classification model to identify the normal customer load profile.

4.4.2 SVM Development

After load profile inspection, a new data set created with 643 normal profiles and 59 abnormal profile with total 702 in which represent 72% of the overall metered electricity dataset from Ruiru residential to enter the model development phase and used to train the SVM classifier in order to create an SVM model. The SVM model development consists of a few stages, which includes: weight adjustment, parameter optimization, and SVM training and testing.

4.4.3 Weight Adjustment

The ratio between the two classes of samples is still unbalanced, class one (no fraud) having 643 samples and class 2 (with fraud) having 59 samples; therefore, SVM classifier technique has a parameter set used to weight and balance the samples' ratio. The weights are adjusted by calculating the sample ratio for each class. This is achieved by dividing the total number of classifier samples with the individual class samples. In addition, class weights are also multiplied by a weight factor of 100 in order to achieve satisfactory weight ratios for training. Fig 4.11 as in rapid miner.

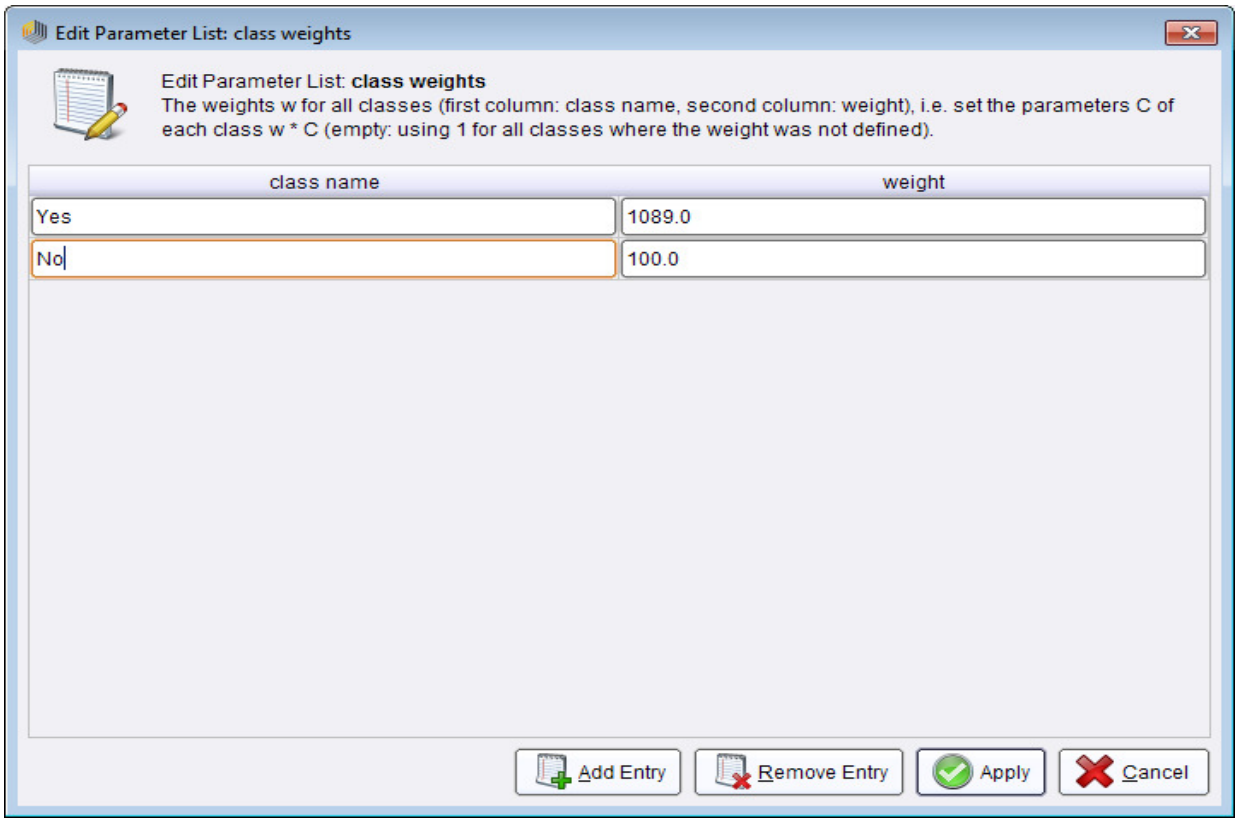


Figure 4.1.3 SVM Label classes weighted as 1089% for YES and 100% for NO.

The other techniques (ANN and KNN) do not have a parameter to balance class labels, so the under-sampling technique was applied to balance classes, under-sampling is a popular method in addressing the class imbalance problem, which eliminates training examples of the over-sized class until it matches the size of the other class. Since it discards potentially useful training examples, the performance of the resulting classifier may be degraded.

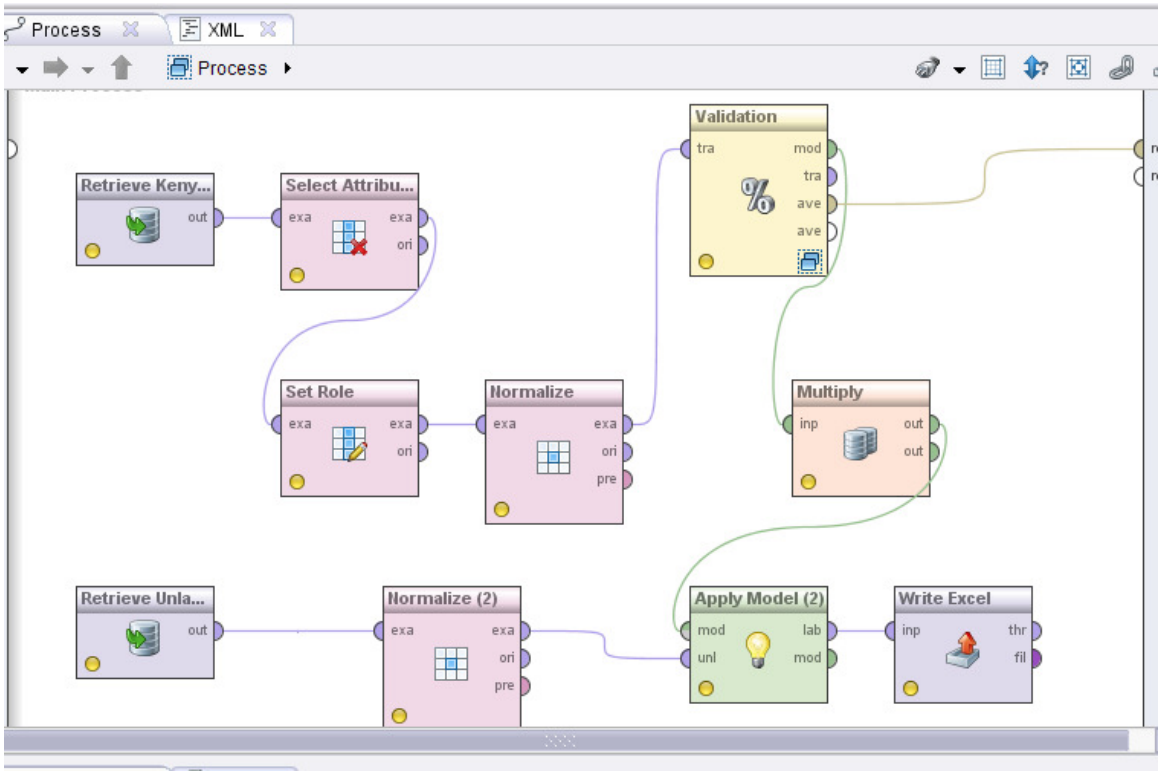
4.4.4 Parameter Optimization

To find the optimal values of the SVM parameters, three parameters were defined and entered as inputs to the grid search technique. The Grid Search method proposed by Hsu et al. in (Chang and C.-J. Lin, 2003) is used for SVM parameter optimization. In the grid search method, exponentially growing sequences of parameters (C, Gamma) are used to identify SVM parameters obtaining the best 10-fold CV accuracy (Jawad Nagi, 2009). The maximum performance achieved when c and gamma parameters equal zeros.

4.4.5 SVM Training and Testing

After replacing missing attributes and normalization, all 702 samples are trained in order to build an SVM model (classifier). In order to facilitate fair comparison between classifiers, the same number of model inputs was adopted for the experiment. Rapid Miner software version 5.3.015 was used to train and test the classifier as shown in the figure 4.4.5 below.

In order to classify customers as abnormal or normal, ten-fold Cross validation was used to test and evaluate four classifiers which are SVM, Neural Network, K-NN and Naïve bayes. Each classifier trained and tested on consumptions profile features.



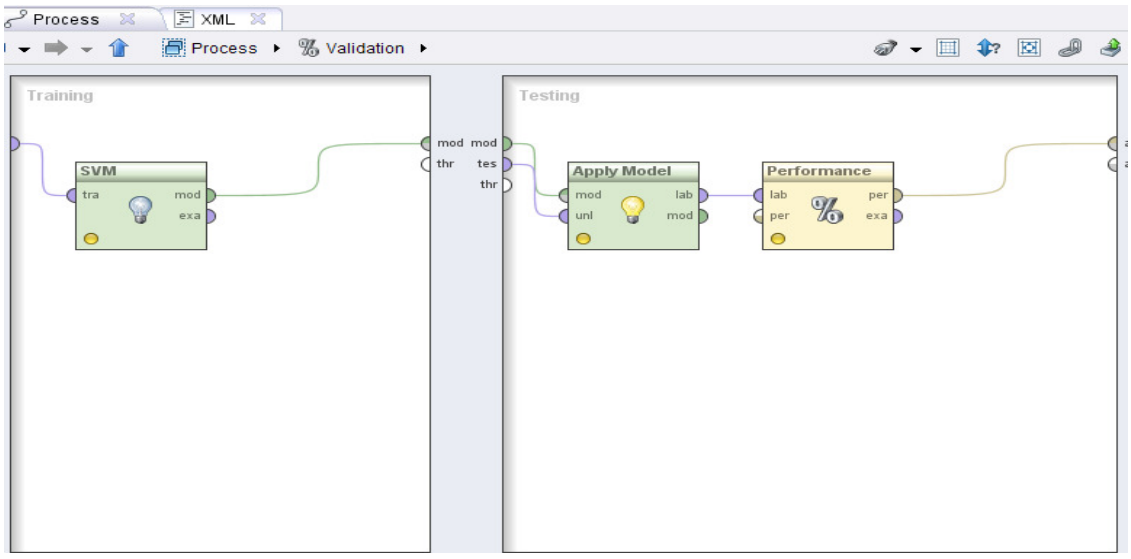


Figure 4.4.5 SVM Classifier training and testing using cross validation

CHAPTER FIVE

5.0 DATA ANALYSIS, MODEL RESULTS AND DISCUSSION

5.1 Introduction

This chapter presents data obtained from the respondents, the analysis, model development, results and discussion. The aim of this study was to develop an automated strategy to preprocess customer data for smoothening noise, feature extraction, data normalization and classification, to explore and identify strategies used by Kenya power to reduce non technical loss of commercial power and to develop a fraud detection model and compare their accuracies using Artificial Intelligence technique such as Support Vector Machine (SVM), Neural Network, K- Nearest neighbor and Naïve bayes. The section also presents the evaluation results between the four selected classification techniques.

5.2 Causes of Non- Technical Electricity Losses

From a majority of the respondents, there are a number of causes of electricity losses and they include: the high cost of electricity that makes it unaffordable, nonpayment of bills of the already used units, poverty levels and unemployment that lead to vandalism of transformers and theft of copper conductors for sale, and competition in manufacturing and production industries which makes some to minimize their production costs through power theft. There is also the case of some customers reconnecting supplies before clearing their bills and paying the reconnection fees.

The rest of the respondents revealed that the lack of adequate mechanism to monitor energy sales vis-à-vis energy dispatches, delayed replacement of faulty meters, wrong allocation of meters on site, delays in entering them in the billing system, and delay in metering new customers upon connection to the service lines contribute largely to non-technical electricity losses.

Further, a few of the respondents said that falsification of readings, meter tampering, direct connection (hooking up), wrong meter reading, faulty and defective metering practices, poor earthing methods at the metering point, and wrong tariff allocation form the other causes of non-technical electricity losses. Integrity and ethics in terms of corruption among Kenya Power staff,

time wasting by staff leading to reduced out-put, wrong billing due to wrong reading of meters, reconnection fees and rebilling, and professional negligence also contribute to non-technical electricity losses.

The above findings reveal that poverty and the high costs of electricity are the major causes of electricity theft which in turn contribute largely to non-technical electricity losses. Poverty due to unemployment makes people to steal electricity since they cannot afford but they need it. They also engage in acts of vandalism where they sell electricity equipments like oil and copper found in the transformers. High electricity costs means that customers end up not paying their bills in time or forfeiting the payment completely leading to huge amounts of uncollected revenues by Kenya power. Since they still need power, they engage in meter tampering, reconnecting illegally, and compromising Kenya Power staff to aid them in all these acts including falsification of meter readings to reduce the billed amount.

5.3 Classifiers performance

The preprocess data trained using the following Artificial Intelligence technique; Support Vector Machine (SVM), Neural Network, K- Nearest neighbor and Naïve bayes. The unlabeled data was tested and the output file generated classifying customers as either normal customers or customers found with anomaly in their electricity consumption. Each data set has been trained and tested per each classifier and results recorded as shown below.

5.3.1 SVM Classifier Confusion Matrix

	True No	True Yes	class precision
Pred No	643	8	98.77%
Pred Yes	0	51	100.00%
Class recall	100.00%	86.44%	
Accuracy	**98.86%**		
Classification error	**1.14%**		

Table 5.3.1 SVM classifier confusion matrix

From table 5.3.1 the results denote that SVM classifier gets 98.86 best accuracy score when dealing with electricity consumptions profile attributes. The classifier classified all normal customers correctly while out of 59 customers found with anomaly 8 customers were misclassified as normal customers thus giving a class recall of 86.44%. The probability of misclassification is approximately 1.14% as given by classification error. The class precision is 98.77% for prediction 'No' and 100% for prediction 'Yes'.

5.3.2 Neural Network Classifier Confusion Matrix

	True No	True Yes	Class Precision
Pred No	635	11	98.30%
Pred Yes	8	48	85.71%
Class recall	98.76%	81.36%	
Accuracy	**97.29%**		
Classification error	**2.71%**		

Table 5.3.2 Neural Network classifier confusion matrix

From table 5.3.2 the results denote that the perception classifier gets 97.29 best accuracy score when dealing with electricity consumptions profile attributes. The classifier classified 635 normal customers correctly while 8 customers were misclassified as customers with anomaly. Out of 59 customers found with anomaly 11 customers were misclassified as normal customers thus giving a class recall of 81.36%.The the probability of misclassification is approximately 2.71% as given by classification error. The class precision is 98.30% for prediction 'No' and 85.71% for prediction 'Yes'.

5.3.3 K-Nearest neighbor classifier confusion matrix

	True No	True Yes	Class Precision
Pred No	643	9	98.62%
Pred Yes	0	50	100.00%
Class recall	100.00%	84.75%	
Accuracy	**98.72%**		
Classification error	**1.28%**		

Table 5.3.3 K-Nearest Neighbor classifier confusion matrix

From table 5.3.3 the results denote that K-Nearest Neighbor classifier gets 98.72 best accuracy score when dealing with electricity consumptions profile attributes. The classifier classified all normal customers correctly while out of 59 customers found with anomaly 9 customers were misclassified as normal customers thus giving a class recall of 84.75%.The the probability of misclassification is approximately 1.28% as given by classification error. The class precision is 98.62% for prediction 'No' and 100% for prediction 'Yes'.

5.3.4 Naïve bayes classifier confusion matrix

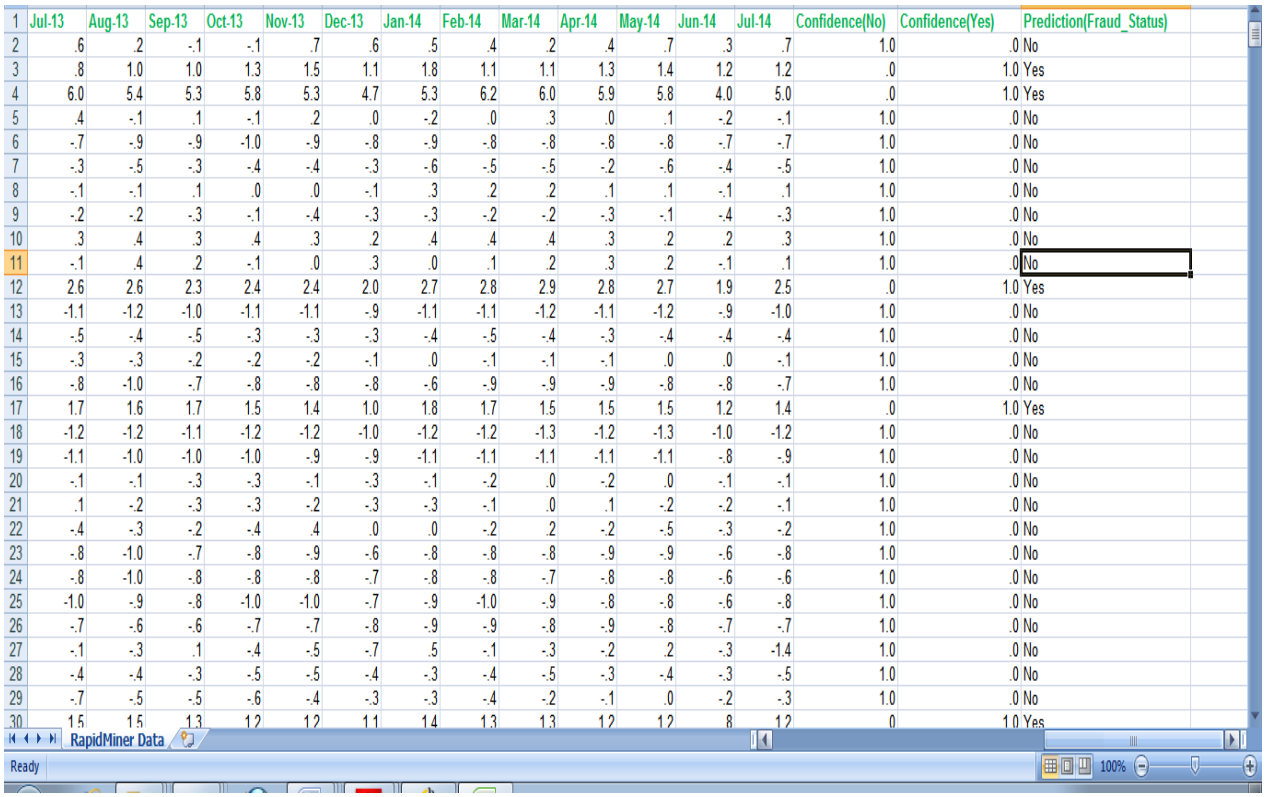
	True No	True Yes	Class Precision
Pred No	582	15	97.49%
Pred Yes	61	44	41.90%
Class recall	90.51%	74.58%	
Accuracy	**89.18%**		
Classification error	**10.84%**		

Table 5.3.4 Naïve bayes classifier confusion matrix

From table 5.3.4 the results denote that the naïve bayes classifier gets 89.18 best accuracy score when dealing with electricity consumptions profile attributes. The classifier classified 582 normal customers correctly while 61 customers were misclassified as customers with anomaly. Out of 59 customers found with anomaly 15 customers were misclassified as normal customers thus giving a class recall of 74.58%.The the probability of misclassification is approximately 10.84% as given by classification error. The class precision is 97.49% for prediction 'No' and 41.90% for prediction 'Yes'.

5.3.5 Deployment on unlabeled instances

After the unlabeled instance on electricity consumption was tested through the classifiers mentioned above, the generated data file (prediction.xls) can be visualized using a spreadsheet application as shown below in fig 5.3.5. It is observed that: the prediction column (Fraud status); and the confidence (the estimated posterior probabilities) for each value of the class attribute. These values (posterior probabilities) are valuable to evaluate the reliability of the prediction.



	Jul-13	Aug-13	Sep-13	Oct-13	Nov-13	Dec-13	Jan-14	Feb-14	Mar-14	Apr-14	May-14	Jun-14	Jul-14	Confidence(No)	Confidence(Yes)	Prediction(Fraud_Status)
2	.6	.2	-.1	-.1	.7	.6	.5	.4	.2	.4	.7	.3	.7	1.0	.0	No
3	.8	1.0	1.0	1.3	1.5	1.1	1.8	1.1	1.1	1.3	1.4	1.2	1.2	.0	1.0	Yes
4	6.0	5.4	5.3	5.8	5.3	4.7	5.3	6.2	6.0	5.9	5.8	4.0	5.0	.0	1.0	Yes
5	.4	-.1	.1	-.1	.2	.0	-.2	.0	.3	.0	.1	-.2	-.1	1.0	.0	No
6	-.7	-.9	-.9	-1.0	-.9	-.8	-.9	-.8	-.8	-.8	-.8	-.7	-.7	1.0	.0	No
7	-.3	-.5	-.3	-.4	-.4	-.3	-.6	-.5	-.5	-.2	-.6	-.4	-.5	1.0	.0	No
8	-.1	-.1	.1	.0	.0	-.1	.3	.2	.2	.1	.1	-.1	.1	1.0	.0	No
9	-.2	-.2	-.3	-.1	-.4	-.3	-.3	-.2	-.2	-.3	-.1	-.4	-.3	1.0	.0	No
10	.3	.4	.3	.4	.3	.2	.4	.4	.4	.3	.2	.2	.3	1.0	.0	No
11	-.1	.4	.2	-.1	.0	.3	.0	.1	.2	.3	.2	-.1	.1	1.0	.0	No
12	2.6	2.6	2.3	2.4	2.4	2.0	2.7	2.8	2.9	2.8	2.7	1.9	2.5	.0	1.0	Yes
13	-1.1	-1.2	-1.0	-1.1	-1.1	-.9	-1.1	-1.1	-1.2	-1.1	-1.2	-.9	-1.0	1.0	.0	No
14	-.5	-.4	-.5	-.3	-.3	-.3	-.4	-.5	-.4	-.3	-.4	-.4	-.4	1.0	.0	No
15	-.3	-.3	-.2	-.2	-.2	-.1	.0	-.1	-.1	-.1	.0	.0	-.1	1.0	.0	No
16	-.8	-1.0	-.7	-.8	-.8	-.8	-.6	-.9	-.9	-.9	-.8	-.8	-.7	1.0	.0	No
17	1.7	1.6	1.7	1.5	1.4	1.0	1.8	1.7	1.5	1.5	1.5	1.2	1.4	.0	1.0	Yes
18	-1.2	-1.2	-1.1	-1.2	-1.2	-1.0	-1.2	-1.2	-1.3	-1.2	-1.3	-1.0	-1.2	1.0	.0	No
19	-1.1	-1.0	-1.0	-1.0	-.9	-.9	-1.1	-1.1	-1.1	-1.1	-1.1	-.8	-.9	1.0	.0	No
20	-.1	-.1	-.3	-.3	-.1	-.3	-.1	-.2	.0	-.2	.0	-.1	-.1	1.0	.0	No
21	.1	-.2	-.3	-.3	-.2	-.3	-.3	-.1	.0	.1	-.2	-.2	-.1	1.0	.0	No
22	-.4	-.3	-.2	-.4	.4	.0	.0	-.2	.2	-.2	-.5	-.3	-.2	1.0	.0	No
23	-.8	-1.0	-.7	-.8	-.9	-.6	-.8	-.8	-.8	-.9	-.9	-.6	-.8	1.0	.0	No
24	-.8	-1.0	-.8	-.8	-.8	-.7	-.8	-.8	-.7	-.8	-.8	-.6	-.6	1.0	.0	No
25	-1.0	-.9	-.8	-1.0	-1.0	-.7	-.9	-1.0	-.9	-.8	-.8	-.6	-.8	1.0	.0	No
26	-.7	-.6	-.6	-.7	-.7	-.8	-.9	-.9	-.8	-.9	-.8	-.7	-.7	1.0	.0	No
27	-.1	-.3	.1	-.4	-.5	-.7	.5	-.1	-.3	-.2	.2	-.3	-1.4	1.0	.0	No
28	-.4	-.4	-.3	-.5	-.5	-.4	-.3	-.4	-.5	-.3	-.4	-.3	-.5	1.0	.0	No
29	-.7	-.5	-.5	-.6	-.4	-.3	-.3	-.4	-.2	-.1	.0	-.2	-.3	1.0	.0	No
30	1.5	1.5	1.3	1.2	1.2	1.1	1.4	1.3	1.3	1.2	1.2	.8	1.2	0	1.0	Yes

RapidMiner Data

Ready

100%

Figure 5.3.5 the generated data file (prediction.xls)

5.4 Classifiers Comparison results analysis

The trained and tested dataset denote that SVM classifier get the best accuracy of 98.86%, followed by K-Nearest neighbor at 98.72, Neural network at 97.29% and the least with lowest score accuracy being naïve bayes at 89.18%. The biggest score for the fraud detection hit rate is achieved by SVM classifier which is 86.44% and the next classifier with optimal fraud detection rate is the K-Nearest Neighbor at 84.75%.The SVM classifier stand out being best with least classification error of 1.14% while the classifier with the highest probability of misclassification is the Naïve bayes with 10.84%.

The results presented in tables 5.3.1, 5.3.2, 5.3.3 and 5.3.4 imply that SVM got the best results in all balanced datasets. The SVM predict the instance class as a black box namely without induce a descriptive rule explains the attributes that define the class label, this research concentrate on determining if the tuple fraudulent or not in addition to the SVM classifier has a notable number of advantages when compared with other classifiers such as Neural Network, Naïve bayes and K-Nearest Neighbor. Firstly, SVM has non-linear dividing hypersurfaces that give it high discrimination. Secondly, SVM provides good generalization ability for unseen data classification. Thirdly, apart from performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. In addition, SVM determines the optimal network structure itself, which is not the case with traditional neural networks. With the introduction of the applied SVM SVC technique, the developed model can control the balance between sensitivity and specificity, giving the fraud detection system more flexibility. Thus, for this type of research study, SVM is more practical and favorable, where this control is most needed in the frequent presence of unknown and unbalanced data sets as the researcher done in (Jawad Nagi, 2009).

CHAPTER SIX

6.0 CONCLUSION AND FUTURE WORK

This chapter concludes the research and summarizes the research contributions made. The achievements and objectives of the research study with respect to the project are highlighted along with the key findings of the research. In addition, this chapter also discusses the impact and significance of this research project to the Kenya Power and suggests possible future research areas.

In this research study, the SVM, Neural Network, Naive bayes and K-Nearest techniques have been investigated and applied to the development of the proposed fraud detection model. The main concern of this research study is the application of SVM for the classification of patterns on load consumptions profiles for customers into two categories: normal (No) and fraud (Yes) customers.

From the result analysis of the tested models, the SVM model emerged as the best model in comparisons with other artificial intelligence techniques. The researcher is therefore proposing the SVM model to be adopted by the Kenya Power for the classification of customers into either normal customers or customers with anomaly based on their monthly power consumption. The SVM provides the use of soft margins for the purpose of separation (classification), thus allowing improvement in the generalization performance of the system.

The training accuracy achieved by the SVC is 98.86%, which indicates that the memorization and learning capability of the SVC model is notably good, even with the presence of noisy data. The main reason which contributes to this good training accuracy is proper selection of the features for building the training model and fine tuning of the SVC and RBF kernel parameters. Thus, the fraud detection model using SVC model is able to achieve an approximate inspection hitrate of 86.44% which increases Kenya Power current inspection hitrate which mostly depend on manual inspection or depends on report from the public.

Though the other classifiers models were analyzed and tested alongside SVM, this study adopted the SVM classifier for the following reasons. First, balancing technique used for ANN, KNN and K-Nearest Neighbor depend on random sampling which decrease the number of instances in the

training data set to more than the half. Second, the SVM classifier depends on class weighting technique to balance data set without omitting any instance and finally SVM got the maximum accuracy score with balanced data sets.

6.1 Achievement of the research objectives

The main objective of this research is to develop a fraud detection model in order to assist Kenya power company fraud control team to increase effectiveness of their onsite operation for reduction of NTLs. Through this research study the desired fraud detection model was developed successfully and the researcher believe that it will be able to help fraud control team of Kenya Power identify and detect customers with fraud activities.

The second objective of this research aims to develop an automated strategy to preprocess customer data for smoothening noise, feature extraction, data normalization and classification. The data preprocessing framework used to implement the fraud detection system is illustrated in Figure 4.2

6.2 Research Contribution

The possible reason as to why customer installation inspections are carried out at random, is due to unavailability of a detection system that can shortlist possible suspicious customers with fraud activities. With the use of a fraud detection model, the customer installation inspections are guided to a reduced group of likely fraudulent customers thus higher fraud detection rate of success is achieved. In addition, the proposed fraud detection model will provide Kenya Power with information tool for efficient detection and classification of NTL activities, in order to increase effectiveness of their onsite inspection operations. With the implementation of the proposed model, operational costs for Kenya Power due to onsite inspection in monitoring NTL activities will be significantly reduced. This will also reduce the number of inspections carried out at random, resulting in a higher fraud detection hitrate and finally knowledge dissemination and behavior regarding fraudulent consumption patterns is obtained by the use of the proposed model, which is useful for further study and analysis by NTL experts and inspection teams in Kenya Power.

6.3 Future Work.

The research study focused on Kenya power customers residing within Ruiru area with only about 1000 customers taken into consideration. There is great need to expand and by classifying load consumption patterns across the country. This approach seemed to be more accurate, as different regions might have different consumption patterns and trends and with large data sample. Moreover, customer power consumption irregularities keep on increasing, thus the need to collect more customer irregularities data in order to make bigger training data set for further testing and evaluation to increase the fraud detection hit rate and improve the proposed model accuracy with more customer data. Finally there is need to integrate the proposed fraud detection model with a user-friendly Graphical User Interface (GUI).

REFERENCE

AihuaShen, R. Tong (2007) “Application of classification Models on credit card Fraud Detection”.

Babbie, E. (2002). *The Basics of Social Research*. Belmont, CA: Wadsworth Publishing.

C.-W. Hsu, C.-C. Chang and C.-J. Lin. (2003.) “*A Practical Guide to Support Vector Classification*”, technical Report, department of computer science and Information engineering, national taiwan university, Taipei.

Daniel Morariu, Lucian N. Vintan, and Volker Tresp. (2006) “*Evaluating some Feature Selection Methods for an Improved SVM Classifier* ", International Journal of Intelligent Technology Volume 1 Number 4.

David L. Olson DursunDelen "*Advanced Data Mining Techniques*".

Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li. (2007) "*A Review of Data Mining-based Financial Fraud Detection Research*".

Dick, A.J.(1995) *Theft of Electricity—how UK Electricity companies detect and deter*. In: Proceedings of the European Convention Security and Detection, Brighton, UK.

Glenn J. Myatt. (2007) “*Making Sense of Data A Practical Guide to Exploratory Data analysis and Data Mining*”.

Glenn J. Myatt. (2007) “*Making Sense of Data A Practical Guide to Exploratory Data analysis and Data Mining*”.

http://en.wikipedia.org/wiki/*Support_vector_machine*, (2014, October).

http://www.statsoft.com/textbook/statistics-glossary/l/?button=0#LinearModeling (online text book, 2014 December).

Jawad Nagi, Keem Siah Yap, Sieh Kiong Tiong, Member, IEEE, Syed Khaleel Ahmed, Member, IEEE and Malik Mohamad.(2010) "*Non technical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines*".

Jawad Nagi. (2009) "An *intelligent system for detection of non-technical losses inTenaga Nasional Berhad (TNB) malaysia low voltage distribution network*", master's dissertation, university tenaga national, Malaysia.

Jiawei Han and Micheline Kamber. (2005)" *Data Mining: Concepts and Techniques Second Edition*".

Konstantinos Tsiptsis, AntoniosChorianopoulos. (2009)"*Data Mining Techniques In CRM*".

Kothari, C.R. (2004). *Research Methodology-Methods and Techniques*, (2nd Edition), New Age International Publishers.

KPLC Annual report (2010). *Annual report and financial statements for the year ended 30th June 2010.*

M.A. Maloof. (2003) " *Learning when datasets are imbalanced and when costs are unequal and unknown*,"in Working Notes of the ICML'03Workshop On Learning from Imbalanced DataSets, Washington, DC.

Nongye. (2003) "*The handbook of data mining* ".

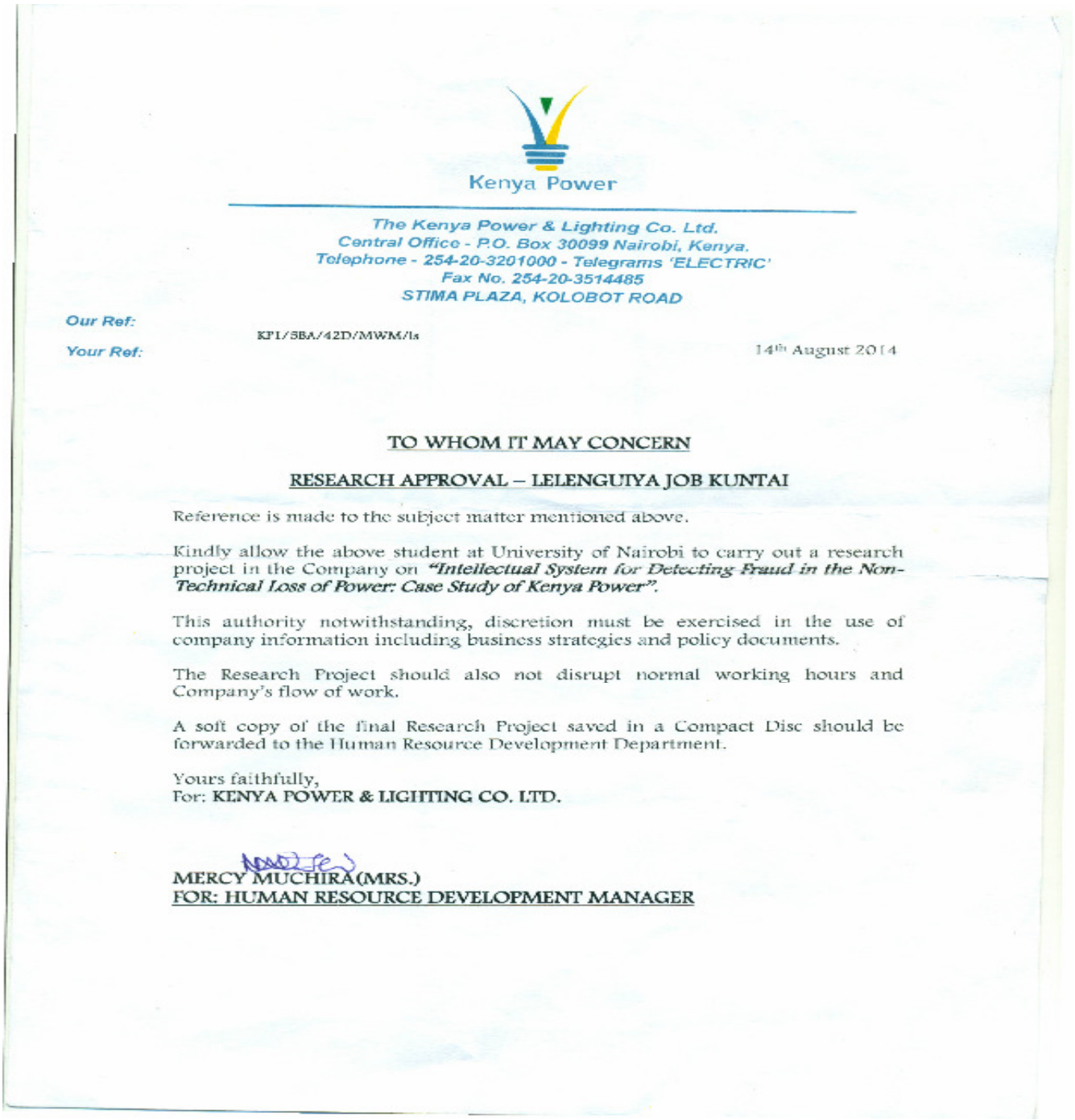
Oded Z. Maimon, LiorRokach. (2008) "*Data Mining with Decision Trees: Theroy and Applications*".

Rapid Miner 5.3, http://www.rapidminer.com, (2014, November).

Smith, T.B. (2003). *Electricity theft: A comparative analysis,* Energy Policy, Vol. 32, pp. 2067-2076.

APPENDIX

Research approval letter



Kenya Power

The Kenya Power & Lighting Co. Ltd.
Central Office - P.O. Box 30099 Nairobi, Kenya.
Telephone - 254-20-3201000 - Telegrams 'ELECTRIC'
Fax No. 254-20-3514485
STIMA PLAZA, KOLOBOT ROAD

Our Ref: KP1/5BA/42D/MWM/ls

Your Ref:

14th August 2014

TO WHOM IT MAY CONCERN

RESEARCH APPROVAL – LELENGUIYA JOB KUNTAI

Reference is made to the subject matter mentioned above.

Kindly allow the above student at University of Nairobi to carry out a research project in the Company on *"Intellectual System for Detecting Fraud in the Non-Technical Loss of Power: Case Study of Kenya Power"*.

This authority notwithstanding, discretion must be exercised in the use of company information including business strategies and policy documents.

The Research Project should also not disrupt normal working hours and Company's flow of work.

A soft copy of the final Research Project saved in a Compact Disc should be forwarded to the Human Resource Development Department.

Yours faithfully,
For: KENYA POWER & LIGHTING CO. LTD.

MERCY MUCHIRA (MRS.)
FOR: HUMAN RESOURCE DEVELOPMENT MANAGER